虚拟健康社区文本数据知识发现策略与模型*

■ 牟冬梅[1] 琚沅红[1] 戴文浩[1] 黄丽丽[2]

[1] 吉林大学公共卫生学院 长春 130000 [2] 长春中医药大学现代教育技术中心 长春 130000

摘要:[**目的/意义**]分析并提出虚拟健康社区文本数据的知识发现策略,构建虚拟健康社区文本数据知识发现模型。[**方法/过程**]通过总结分析虚拟健康社区文本数据特点,针对其特点带来的数据挖掘困难制定相应的知识发现策略,并在 DIKW 体系指导下,依据提出的知识发现策略构建虚拟健康社区文本数据知识发现模型。通过应用计算机编码、自然语言处理技术、句法分析、制定推理规则等方法实现从自由文本数据到药物不良反应智慧的数据价值升华过程。[**结果/结论**]通过实证研究验证提出的知识发现策略和知识发现模型的有效性和可操作性,为后续虚拟健康社区文本数据知识发现的相关理论与实证研究提供参考。

关键词:虚拟健康社区 文本数据 知识发现 知识发现策略 知识发现模型

分类号:G251

DOI:10. 13266/j. issn. 0252 -3116. 2018. 05. 014

近年来,随着互联网的发展和社交媒体的盛行,人们越来越多地选择通过社交媒体获得健康相关信息,虚拟健康社区作为常见的社交媒体类型,逐渐成为互联网中最热门的讨论话题之一,同时也日益引起学术界的广泛关注。虚拟健康社区为用户分享彼此的经验和观点提供了可交互平台,虚拟健康社区中由用户生成的数据量呈爆炸式增长,且虚拟健康社区数据中蕴含着大量有价值的信息,这为知识发现提供了新的研究领域。面对海量的、文本化的、文本内容表述不规范的虚拟健康社区数据信息,如何从中发现领域用户感兴趣的领域相关信息是当下面临的一大挑战。虚拟健康社区作为一种新兴的领域研究数据来源,相比传统文献数据库与科研实验数据库,其特色在于:①虚拟健康社区具备海量的数据资源可供挖掘使用;②虚拟健康社区中的领域数据更加贴近用户的真实情况的反应;③虚拟健康社区中的数据由用户自愿生成;④虚拟健康社区中的数据具有更好的时效性,数据更新速度也更快。从以上描述可以看出,虚拟健康社区具有数据传输速度快、应用范围广、更新频率快等特征,且其中蕴含大量数据形式复杂多样、价值深埋有待挖掘的虚拟健康社区数据,为数据挖掘及知识发现奠定了坚实的数据基础,但是虚拟健康社区数据的潜在有价值信息需要通过进行处理才能获得其中隐含的知识和智慧。DIKW(data-information-knowledge-wisdom)体系呈现了从数据到信息、再到知识的层层沉淀凝练最终到智慧的转化过程。因此基于 DIKW 体系从数据 - 信息 - 知识 - 智慧的转换过程可抽象出一个通用方法模型,为领域用户对虚拟健康社区文本数据进行知识发现研究提供指导。

1 基于社交媒体的健康信息知识发现研究现状

虚拟健康社区,作为较为典型的具有学科领域特色的社交媒体中的一种,随着人们越来越多地到虚拟健康社区中分享和寻找健康相关信息而成为值得重点关注的研究对象。通过对社交媒体挖掘[1]定义的理解,虚拟健康社区挖掘可以被视为一个从虚拟健康社区数据中表示、分析和提取可操作的模式的过程。虚拟健康社区中数据的飞速增长、数据挖掘技术的发展和生物医学领域工具和资源的积累,使得从虚拟健康社区数据中发现潜在的有用知识的研究有了可靠的信息源和技术支撑。

* 本文系国家自然科学基金项目"嵌入式知识服务驱动下的领域多维知识库构建"(项目编号:71573102)和吉林省教育厅社科项目"虚拟健康社区知识发现与实证研究"(项目编号:JJKH20170881SK)研究成果之一。

作者简介:牟冬梅(ORCID:0000 -0003 -0237 -034x),教授,博士生导师;琚沅红(ORCID:0000 -0002 -9146 -4788),硕士研究生;戴文浩(ORCID:0000 -0002 -5796 -1342),硕士研究生;黄丽丽(ORCID:0000 -0003 -2651 -8089),讲师,通讯作者,E-mail:huanglili1218@126. com。

收稿日期:2017 -09 -10 修回日期:2017 -10 -20 本文起止页码:125 -131 本文责任编辑:王传清

目前基于社交媒体的健康信息挖掘的研究主要集中在热点主题识别和疾病趋势预测。X. Ji 等[5]使用两步情感分析法从 Twitter 消息的情感分类来衡量用户对某种疾病的关注等级，预测在疾病影响下用户所标识的关注程度；D. Ghosh 等[6]以肥胖为例，使用 LDA (latent dirichlet allocation) 主题模型和空间分析方法以识别 Twitter 中健康相关的话题；R. Mehrotra[7]提供了两种新的方法改进 LDA 主题模型以对微博内容行识别，该研究为用户提供了一种新的方法，可以显著改建 LDA 主题建模，而无需对底层 LDA 机制进行修改。J. Parker 等[8]提出一个通用框架以探测 Twitter 中爆发的公共健康趋势，该研究局限性在于使用 Wikipedia 和 ICD 只能探测到之前已知的疾病，无法探测到新疾病。S. Doan 等[9]提出了一种新的过滤方法，从587 000 000 条 Tweet 中筛选流感样疾病（influenza-like-ilnesses, ILI）。研究中使用的语义特征过滤的方法对于基于地理选择 tweets 是非常有用的，可以广泛地适用于疾病和综合症状，如可通过限定某个或某些区域来探测该区域流感样疾病的情况。P. Kostkova 等[10]通过探测疾病的传播来展示社交网络的早期预警的能力，并展示了在猪流感期间在线资源是如何传播的，研究显示社交媒体的实时更新性能够用来完善疫情早期预警探测系统，但是对于位置的识别仍是一个问题。S D. Young 等[11]使用社交媒体实时技术探测和远程监视 HIV 的暴发，该研究受到许多因素的影响，如缺乏及时的艾滋病病例报告，不能实时评估最近的艾滋病病毒感染或艾滋病危险行为之间的关系。由此产生的 HIV 和艾滋病病毒相关的 Tweets 之间的关系在评估时，其参与者的生活地区已是艾滋病毒流行的地区。D. Barazanji 等[12]提出一个系统来实现点源爆发的监测，该研究主要是在已知疾病基础上的疾病暴发监测，无法对之前未发生过的疾病进行预测。综上所述，领域用户需要来自网络的信息，尤其是虚拟健康社区的数据。而目前对于虚拟健康社区知识发现的研究主要集中在通过对虚拟健康社区数据内容的分析发现领域的研究热点、研究前沿和研究趋势，且在技术方面，大多是关于信息抽取方法和技术的应用研究，鲜有将语言学理论、自然语言处理技术、学科领域知识及数据挖掘理论相结合，进而指导虚拟健康社区文本数据中隐含知识发现的研究。因而，本研究将结合自然语言处理技术、句法分析、主题模型、本体映射等理论技术与方法应用于虚拟健康社区文本数据知识发现的研究中。

2 虚拟健康社区文本数据知识发现策略

Twitter、Facebook、微博等大众社交媒体和 DailyStrength、Medhelp 等虚拟健康社区平台中每天都会生成大量的健康相关数据，这些由用户生成的、数量众多的与疾病诊断、药物进展及药物不良反应等健康相关数据有非常重要的研究和应用价值。与文献数据、科研数据相比，虚拟健康社区文本数据不规范，主要表现在：数据形式是无结构的自由文本；概念描述用词口语化、习惯用语程度高、存在大量字符缺失、单复数混用等现象；实体语义关系通过语境来体现，并未给予直观的抽象；虚拟健康社区数据是表达个人感受的平台，客观事件夹杂着情感表达，使得事件陈述更加模糊；大量的知识隐含在事件中也未显现。针对这些困难，在语言学、信息组织、计算机科学等理论的指导下，分别对知识发现过程中的实体识别、语义关系抽取和事件探测问题制定知识发现策略，在此基础上最终形成较为完整的虚拟健康社区数据挖掘与知识发现策略，从而指导虚拟健康社区文本数据中的知识发现问题的分析与解决。

虚拟健康社区知识发策略针对虚拟健康社区文本数据特点产生的原因和各个特点的表现形式进行了总结与分析，且对于虚拟健康社区文本数据的各个特点，结合信息组织理论、本体映射理论，计算机技术、自然语言处理技术等，提出了有针对性的解决方案，从而指导虚拟健康社区文本数据的知识发现研究，见图 1。由于信息交流表现方式和信息存储的自身需求使得虚拟健康社区数据呈现文本化特点，表现为数据的非结构化，且多以自由文本形式存储于网络中。因此，针对文本化特点，可通过构建半结构化文本库存储来自网络的虚拟健康社区文本数据；由于用户生成内容用词不规范、用户自身素养存在的差异性及用户表达方式的随意性导致虚拟健康社区文本数据具有概念描述口语化、关系表达自由化、事件阐述模糊化及知识蕴含隐蔽性特点，表现为用户用词的非专业化、文本中概念关系类型多样、文本内容组织无序化及数据分散和知识表达的模糊化，针对这些特点，可采用句法分析、制定语法规则、关系主题模型和领域本体等来解决。基于以上虚拟健康社区文本数据特点和解决方案，本研究针对虚拟健康社区数据中实体关系表达的自由化提出基于语法规则的实体语义关系抽取策略和针对事件阐述的模糊性提出事件探测策略，通过对虚拟健康社区知识发现策略的研究可以加深对知识发现理论的认知和理解，有助于促进知识发现的应用向纵深方向发展，从而提高虚拟健康社区自由文本中知识发现的速度和效率。

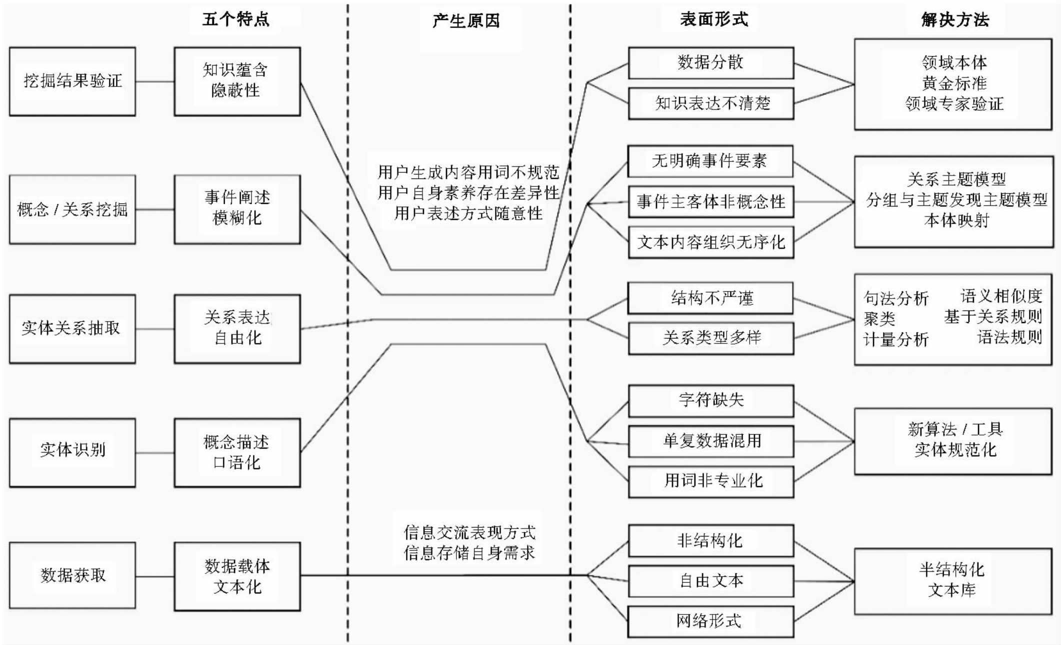


图1 虚拟健康社区数据知识发现策略

3 虚拟健康社区文本数据知识发现模型

在针对虚拟健康社区数据存在的各个特点形成有针对性的知识发现策略，进而明确虚拟健康社区数据挖掘与知识发现方案基础上，需要更多地关注面向虚拟健康社区的文本数据知识发现研究的核心问题，以进一步明晰工作流程，这就需要对虚拟健康社区文本数据知识发现模型进行研究，通过对模型的各个组成要素及要素之间的关系对知识发现策略的科学性和可行性进行验证。

3.1 DIKW 体系

美国管理学家罗素・艾可构建了 DIKW 体系[13]，DIKW 体系是关于数据、信息、知识及智慧的体系，当中每一层对下一层赋予某些特质[14]。DIKW 体系如表1 所示：

表1 DIKW 体系

类别(class)	价值(value)	目的(purpose)	方法技术(method/technology)
数据(data)	原始素材	know-nothing(一无所知)	数据筛选、计算机编码
信息(information)	加工处理后有逻辑的数据	know-what(知道是什么)	数据库技术、句法分析、实体识别
知识(knowledge)	提炼信息之间的联系，行动的能力，完成当下任务	know-how(知道是怎样)	本体映射、关系主题模型
智慧(wisdom)	关心未来，具有预测的能力	know-why(知道是为何)	数据挖掘技术、专家验证

根据 DIKW 理论总结由数据到智慧的知识发现，首先，对网页、虚拟健康社区等领域的异构海量数据进行数据抽取；其次，进行数据筛选、数据清洗，使数据呈现结构化、模型化；再次，进行信息整合、统计分析和综合归纳以形成知识；最后，进行隐性知识挖掘，为用户提供个性化知识服务及辅助决策支持的智慧。由此可见，智慧是从数据到信息、再到知识的层层沉淀凝练所得，其间要经过数据采集、数据结构化、自然语言处理、语义化、事件探测与知识发现等过程。虚拟健康社区文本数据的知识发现恰恰符合 DIKW 体系从数据到智慧的过程。

3.2 DIKW 指导下的虚拟健康社区知识发现模型

在 DIKW 体系指导下，虚拟健康社区文本数据的知识发现应依次经过如下过程：①应用计算机技术和数据库技术从网络中获取虚拟健康社区文本数据形成原始数据的文本库，这些原始数据须进行被加工处理才具有实际意义；②使用自然语言处理技术对获取的数据进行初步分析，形成信息相对集中的信息数据，为用户提供对数据的初步了解；③对信息数据进行句法分析等，进一步获得包含语义关系的知识；④对包含语义关系的知识进行提炼，最终获得智慧数据。基于此，面向虚拟健康社区数据的知识发现模型由底向上分为5 层：数据层、自然语言处理层、语义分析层、关系抽取层、事件探测层。这 5 层由文本库构建、命名实体识别、实体语义关系抽取、事件探测及知识发现的工作流

程和技术路线加以实现。其中文本库构建完成网络数据向本地数据的转换,数据由非结构化转换为半结构化;命名实体识别在自然语言处理的基础上,将名词通过本体映射实现规范化;实体语义关系抽取则建立在句法分析基础上,根据依存距离构建推理规则,依据推理规则抽取命名实体间的语义关系,基于此完成概念和实体间语义关系的抽取;事件探测及知识发现是基于概念/关系对数据中蕴含的事件进行探测,探测到的事件再通过领域内知识库的验证完成事件抽取,最后在领域专家验证后实现知识发现,如图 2 所示:

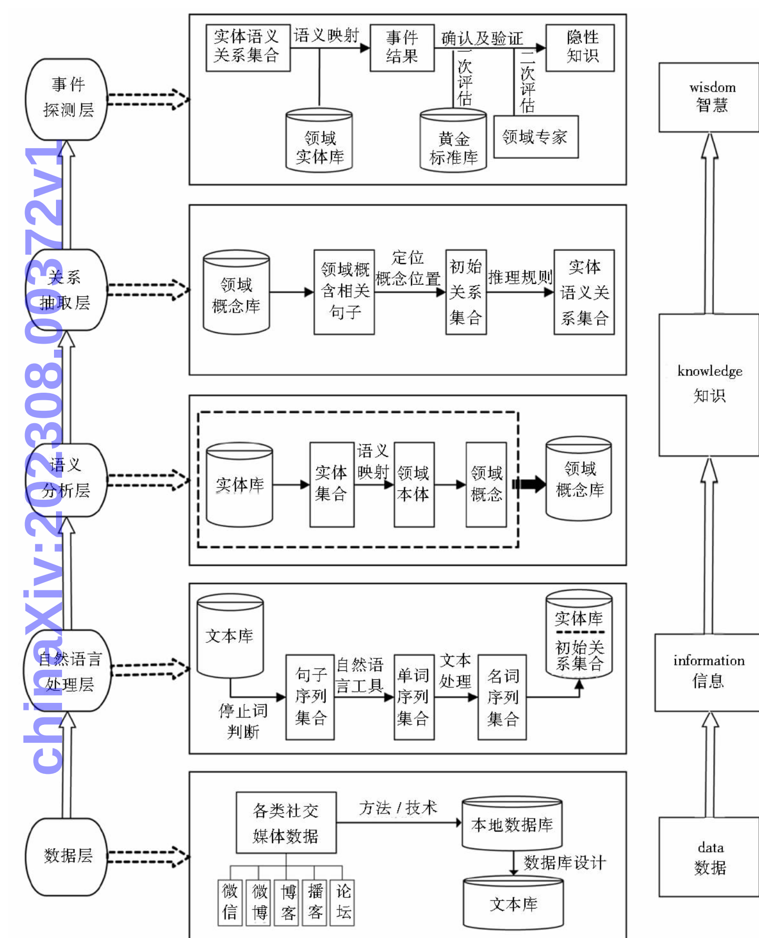


图 2　虚拟健康社区数据知识发现概要模型

在图 2 中,将虚拟健康社区文本数据知识发现模型分为 5 层:数据层、自然语言处理层、语义分析层、关系抽取层和事件探测层。由下而上反映了虚拟健康社区文本数据转化为领域知识与智慧的过程。各层之间的沟通交流通过数据的流动和转换实现。

3.2.1　数据层　数据层是知识发现模型的基础层。主要包括数据源的选择和文本库构建两部分,这一层实现了从虚拟健康社区中目标数据的获取。数据层的功能和任务是为命名实体识别和实体语义关系的抽取作数据准备,是虚拟健康社区数据挖掘与知识发现的起点和保障。

3.2.2　自然语言处理层　主要根据语言学理论实现文本库中的实体识别与抽取和文本句子分析。实体识别与抽取依赖于语言学中词性的分析,选取句子中有实际意义的名词作为抽取的对象,句子分析主要基于语言学中的语法依存理论分析词在文本中的位置。通过这一层的处理,获得包含初始关系的数据集合。

3.2.3　语义分析层　虚拟健康社区文本数据内容的复杂性表现为文本内容中词或概念的使用不规范,且由于表述的自由性,词与词之间的语义关联难以揭示。在语义分析层中,通过借鉴基于顶层本体的语义互联模式解决语义互联的方式,将形式标准化后的命名实体进行概念的规范化,通过使用领域本体与自由文本进行语义映射识别出文本中的领域概念,获得领域概念集合。

wisdom
智慧
knowledge
知识
information
信息
data
数据

3.2.4　关系抽取层　句子的语义由两部分组成:一是组成该句子的词本身的语义;二是句子中词与词之间的语义关系。句子的句法结构和语义关系是实体关系抽取中极其重要的步骤,因此能否正确地抽出文本中实体关系在于是否能够根据句子语义的特性制定出具有较高鉴别能力的抽取实体关系的规则,即推理规则的制定。关系抽取层通过制定推理规则实现实体间语义关系的识别,数据形式由领域概念转换成包含语义关系的概念/关系对。

3.2.5　事件探测层　事件作为信息的重要表达方式,是指特定的人和物在特定时间和特定地点相互作用的一种客观事实[15]。针对虚拟健康社区的自由文本数据,事件探测即通过事件抽取方法从含有事件信息的文本中抽取事件的内容。通过多个领域本体之间的语义映射从实体语义关系识别后的概念/关系对中发掘潜在的事件信息,并与领域知识库或金标准进行对比以发现其中隐含的领域新知识,再经领域专家验证知识发现结果的可靠性和准确性,由这一层实现知识到智慧的提升。

以上是在虚拟健康社区知识发现策略的指导下,以 DIKW 为体系,构建虚拟健康社区文本数据的数据挖掘与知识发现概要模型。下面通过虚拟健康社区药物不良反应的挖掘研究对提出的知识发现策略和构建

牟冬梅，琚沅红，戴文浩，等. 虚拟健康社区文本数据知识发现策略与模型[J]. 图书情报工作,2018,62(5):125－131.

的知识发现模型的可行性进行实证验证。

4 虚拟健康社区中药物不良反应知识发现的实证研究

虚拟健康社区与其他社交媒体不同之处在于其中包含了大量由用户生成的健康相关信息，因此能够反映出用户对于健康疾病诊疗、用药的真实反馈，在此类数据中，药物及其不良反应是人们最为关注的医疗信息类型之一。由于药物发布前临床试验具有时间、试验对象等局限性，不是所有的药物不良反应都能够被识别出来，且药物不良反应造成的后果可能非常严重。此外，在实际临床工作中，临床医生需要了解可能的药物不良反应从而根据患者自身情况调整用药；药品生产厂家也需要掌握药品上市后的实际药物不良反应情况，对药品药效等做进一步的改良。因此，及时、准确地识别药物不良反应对全球公共卫生系统来说都是一个紧迫的、亟需解决的问题。

根据前文所述的 DIKW 体系和依据其构建的虚拟健康社区文本数据知识发现模型可以看到，从虚拟健康社区中药物不良反应的知识发现就是将虚拟健康社区文本数据转化为药物不良反应智慧的过程。本研究中将虚拟健康社区药物不良反应知识发现分为 4 个阶段，药物不良反应数据文本库构建、药物不良反应实体识别与关系抽取、药物不良反应事件探测及药物不良反应事件确认。

4.1 虚拟健康社区药物不良反应数据——数据获取——文本库(data-D)

本研究将在构建的虚拟健康社区文本数据知识发现模型基础上，对虚拟健康社区中潜在的药物不良反应的知识发现进行实证研究。具体是通过虚拟健康社区 MedHelp[16] 中潜在的药物不良反应知识的发现对构建的虚拟健康社区文本数据知识发现模型进行实证研究。研究中以肾脏疾病(kidney disease)版块中的帖子(以下称 posts)作为研究对象，进行数据获取与文本库构建。基于存储信息进行数据库设计，根据 MySQL(mysql-5.6.24-win32)中支持的数据类型，结合 MedHelp 中 posts 的发帖人信息、posts 内容长度等实际情况，设计存储 posts 和对 posts 进行处理的过程中产生的数据的表结构。经过过滤，从 MedHelp 中获取 kidney disease &disorder 主题下药物不良反应相关帖子共计 19929 个。

4.2 药物不良反应信息抽取——自然语言处理——实体/概念库(information-I)

虚拟健康社区文本数据的命名实体识别就是从虚拟健康社区自由文本数据中标注出疾病、药物、症状和副作用等实体[17]。鉴于医学领域本体能够提供医学领域内相关知识的决策支持信息，针对虚拟健康社区数据内容中标准化生物医学词汇与口语化表达共存的特点，通过借助 UMLS[18](一体化医学语言系统)、CHV[19](用户健康词表)、和 SIDER[20](药物不良反应数据库)等医学领域本体对虚拟健康社区自由文本进行语义标注，实现自由文本与领域本体间的语义映射。使用标准化医学知识库 UMLS 及 MetaMap 工具来识别自由文本中的生物医学词汇[21]。

4.3 药物不良反应知识获取——语义关系抽取——语义关系集合(knowledge-K)

根据虚拟健康社区实体语义关系抽取模型的功能描述，首先通过与领域本体映射完成虚拟健康社区文本数据的命名实体识别，识别出自由文本中疾病、症状和副作用概念并抽取出来。接下来，以文本中的句子为一个处理单位，通过制定基于语法分析的推理规则对 post 进行分析，实现虚拟健康社区文本数据中医学领域概念间语义关系的自动分析与抽取，挖掘 posts 中疾病、药物和症状间的语义关系，从而获得药物的不良反应事件信息。研究获得的部分药物不良反应知识发现结果如图 3 所示：

id	mesid	medicalSign	drug	drdis	drsenloc	disease	didis	disenloc	rule
837	4070	kidney issues	prescription medicines	81	1	null	0	0	BA
838	4072	ache	muscle relaxant	22	1	bladder infe	64	-1	BB
839	4072	ache	muscle relaxant	22	1	bladder infe	64	-1	BB
840	4072	ache	muscle relaxant	22	1	bladder infe	64	-1	BB
841	4072	ache	muscle relaxant	22	1	bladder infe	64	-1	BB
842	4072	ache	muscle relaxant	22	1	bladder infe	64	-1	BB
843	4072	ache	muscle relaxant	22	1	bladder infe	64	-1	BB
844	4077	kidney issues	prescription medicines	81	1	null	0	0	BA
845	4153	kidney issues	nsaid	19	0	null	0	0	AA
846	4157	kidney issues	nsaid	41	0	null	0	0	AA
847	4174	kidney issues	nsaid	19	0	null	0	0	AA
848	4178	kidney issues	nsaid	41	0	null	0	0	AA
849	4205	kidney issues	nsaid	19	0	null	0	0	AA
850	4209	kidney issues	nsaid	41	0	null	0	0	AA
851	4226	kidney issues	nsaid	19	0	null	0	0	AA
852	4230	kidney issues	nsaid	41	0	null	0	0	AA
853	4257	tired	sugar	40	0	null	0	0	CA
854	4257	tired	sugar	40	0	null	0	0	CA
855	4293	tired	sugar	40	0	null	0	0	CA
856	4293	tired	sugar	40	0	null	0	0	CA

图 3 实例验证部分结果示意

其中黑色框标记所示为药物 muscle relaxant(肌肉松弛剂)－不良反应 ache(疼痛)关系对。本研究以药物不良反应数据库 SIDER2 中已知的药物不良反应事件信息作为领域金标准，在获得的药物不良反应结果中显示的是通过与 SIDER2 比较，将 SIDER 中已有的 drug-medicalSign 药物不良反应事件对过滤掉，即，只有

经过与 SIDER2 比较后不在其中的 drug-medicalSign 才作为潜在可能的药物不良反应事件呈现给用户。因此研究获得的事件探测结果即为从虚拟健康社区获得的潜在药物不良反应事件信息。

4.4 药物不良反应事件确认——事件探测——隐性知识(wisdom-W)

依据构建的事件探测模型,事件探测需要通过多个领域本体间的语义映射区分文本中哪些语义关系是已经发生的事件,哪些是潜在可能的事件。对通过知识发现模型发现的药物不良反应事件的确认,就是对知识发现结果进行验证和评价以发现隐性知识的过程。因此,研究中所获得的 drug-medicalSign 关系对是否均为药物研究领域的新知识还有待于领域专家和药物、动物以及临床实验的证实。当经过药物、动物实验验证后的药物 - 不良反应关系对,将被补充加入 SIDER 中,成为金标准中的一条记录,从而指导医生临床用药和患者自主服药,并为未来科研工作提供参考,这也是知识发现的最重要的现实意义。

综上,虚拟健康社区中的 posts 经过逐步深入的数据处理与提取,从数据结构上实现了从包含大量噪声的原始自由文本数据向规范化药物不良反应领域知识和智慧的凝练;且对应模型中的各层,每一层中经过处理的信息集合都为更高层的信息分析提供了优质的数据,体现了数据价值的逐步升华,最终为临床用药、患者服药和药物不良事件的减少做出努力和贡献。

5 结语

本研究从虚拟健康社区文本数据挖掘与知识发现的难度和重点入手,针对虚拟健康社区文本数据的非结构化和不规范化特点提出知识发现策略,并在策略的指导下构建虚拟健康社区文本数据知识发现模型,进而基于此理论模型利用美国虚拟健康社区 MedHelp 中数据验证该模型的可操作性,通过制定推理规则的方法实现了对虚拟健康社区中药物不良反应知识的发现。研究中构建的知识发现策略和知识发现模型,将从社交媒体文本数据到其中隐含的领域知识的揭示过程抽象成几个不同分析阶段,通过各个阶段中提出的技术和方法的指导,满足用户对不同层次信息处理的需求,该模型不局限于指导虚拟健康社区的知识发现研究,同样适用于其他领域的知识发现研究。本研究的局限性在于:①研究方法有针对性,虽然本研究的方法在一定程度上能够丰富知识发现的理论和方法,但仍有一些问题需要做进一步研究;②领域本体覆盖面有限,UMLS 超级叙词表作为识别文本中命名实体的领域词典,其本身所包含的概念数量将直接影响 MetaMap 的映射结果,即对 UMLS 之外的词汇的映射效果无法保证;③本体概念不足,针对此问题,未来可通过对领域本体进行更多、更全面的调研,并尝试引入更多领域本体来解决,以便更有效地实现文本中概念的规范化。随着互联网技术和社交媒体技术的急速发展,会有越来越多的理论、技术和方法可用于知识发现模型的构建中。因此在未来的研究中,可能会对提出的基于虚拟健康社区文本数据的知识发现模型进行调整和完善,使其能够更加有效地实现虚拟健康社区中领域新知识的发掘与应用。此外,本研究提出通过制定基于推理规则的数据挖掘方法实现对虚拟健康社区文本数据的数据挖掘与知识发现研究,经验证证实具有可操作性。在未来研究中可针对虚拟健康社区文本数据的特点进行新的数据挖掘与知识发现方法的创新,本研究中构建的知识发现模型可作为新方法的对比模型,以此促进新方法、新技术的开发和应用,进而为用户提供更好的健康知识共享服务。

参考文献:

[1] ZAFARANI R, ABBASI M, LIU H. Social media mining: an introduction[M]. Cambridge: Cambridge University Press,2014:16.
[2] CHEN Y, LI Z, NIE L, et al. A semi-supervised bayesian network model for microblog topic classification[C]// 24th International conference on computational linguistics. Mumbai: COLING, 2012: 561 - 576.
[3] 景悦诚. 基于丰富语言特征的中文社交媒体事件发掘[D]. 上海:上海交通大学, 2015.
[4] 朱晓光. 基于半监督学习的微博情感分析方法研究[D]. 济南:山东财经大学, 2014.
[5] JI X, CHUN S A, GELLER J. Monitoring public health concerns using twitter sentiment classifications[C]// IEEE international conference on healthcare informatics. Philadelphia: IEEE Computer Society, 2013:335 - 344.
[6] GHOSH D, GUHA R. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and geographic information system[J]. Cartography and geographic information science, 2013, 40(2): 90 - 102.
[7] MEHROTRA R, SANNER S, BUNTINE W, et al. Improving LDA topic models for microblogs via tweet pooling and automatic labeling[C]// International ACM SIGIR conference on research and development in information retrieval. Gold Coast: ACM, 2013: 889 - 892.
[8] PARKER J, WEI Y, YATES A, et al. A framework for detecting public health trends with Twitter[C]// IEEE/AMC international conference on advances in social networks analysis and mining.

Niagare Falls:IEEE, 2013:556－563.

[9] DOAN S, OHNO-MACHADO L, COLLIER N. Enhancing twitter data analysis with simple semantic filtering: example in tracking influenza-like illnesses[C]// IEEE second international conference on healthcare informatics, imaging and systems biology. Piscataway:IEEE Computer Society, 2012:62－71.

[10] KOSTKOVA P, SZOMSZOR M, ST LOUIS C. Swineflu: the use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic[J]. ACM transactions on management information systems, 2014, 5(2):1－25.

[11] YOUNG S D, RIVERS C, LEWIS B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes[J]. Preventive medicine, 2014, 63(3):112－115.

[12] BARAZANJI D, BJELKMAR P. System for surveillance and investigation of disease outbreaks[C]//23rd International conference on World Wide Web pages. Seoul: Association for Computing Machinery, 2014:667－668.

[13] ACKOFF R L. From data to wisdom[J]. Journal of applies systems analysis,1989,16(1):3－9.

[14] BELLINGER G, CASTRO D. Data, information, knowledge, and wisdom [J]. Anaesthesia & intensive care medicine, 2004, 15(1):44－45.

[15] 马彬. 事件关系识别关键技术研究[D]. 苏州:苏州大学, 2014.

[16] MedHelp[EB/OL]. [2017－01－25]. http://www.medhelp.org/.

[17] 冯丽芝. 面向命名实体抽取的大规模中医临床病历语料库构建方法研究[D]. 北京:北京交通大学, 2015.

[18] 医学一体化语言系统[EB/OL]. [2016－12－12]. http://www.cintcm.com/yuyan/content/word/UMLS.ppt.

[19] CHV Wiki[EB/OL]. [2017－08－13]. http://consumerhealthvocab.chpc.utah.edu/CHVwiki/.

[20] SIDER[EB/OL]. [2017－09－12]. http://sideeffects.embl.de/.

[21] MetaMap[EB/OL]. [2017－03－12]. http://metamap.nlm.nih.gov/.

作者贡献说明：

牟冬梅：提出研究命题、设计研究思路，对文章关键性内容进行修订、最后审阅论文及定稿，提供资助、支持条件和管理监督；

琚沅红：负责图表整理，论文修改；

戴文浩：负责图表修订；

黄丽丽：负责整理研究思路，文献调研，采集、清洗和分析数据，分析结果、论文撰写。

Knowledge Discovery Strategy and Model of Virtual Health Community Text Data

Mu Dongmei[1] Ju Yuanhong[1] Dai Wenhao[1] Huang Lili[2]

[1] Department of Medical Informatics, School of Public Health, Jilin University, Changchun 130000

[2] Modern Educational Technology Center, Changchun University of Chinese Medcine, Changchun 130000

Abstract: [**Purpose/significance**] This study aims to analyze and propose the knowledge discovery strategy and build a knowledge discovery model of virtual health community text data. [**Method/process**] Firstly it summarized features of virtual health community text data, in view of the difficult of data mining to formulate the corresponding knowledge discovery strategy, and guided by DIKW system, to build knowledge discovery model of virtual health community text data based on knowledge discovery strategy. Through the application of computer code, natural language processing, syntactic analysis, and methods of inference rules, it realized the sublimation process of data value from free text data to the wisdom of adverse drug reactions. [**Result/conclusion**] Empirical research is carried out to verify the effectiveness and operability of the proposed knowledge discovery strategy and knowledge discovery model, so that it can provide reference for the subsequent theory and empirical research on knowledge discovery of virtual health community text data.

Keywords: virtual health community free text data knowledge discovery knowledge discovery strategy knowledge discovery model